nature portfolio

https://doi.org/10.1038/s41586-025-09962-4

Supplementary information

A benchmark of expert-level academic questions to assess AI capabilities

In the format provided by the authors and unedited

Supplementary Information for the HLE Dataset

HLE Team

October 2025

1 Contributor Guidelines

Objective

Together, we are collecting the hardest and broadest set of questions ever. Can you think of something you know that would stump current artificial intelligence (AI) systems? This will help us better evaluate the capabilities of AI systems in the years to come. If you submit a challenging question that passes review, your name will be associated with the question and you will be invited as an author of the paper corresponding to this dataset. Simply think of a hard question, and see if AIs get it right. If it's hard for the AIs it is likely good to submit. If there is an exceptionally difficult question you've given in an exam, or a niche result you've encountered in your research, feel free to make this the basis for a question. We accept challenging questions from all fields, including mathematics, rocket engineering, analytic philosophy, and more.

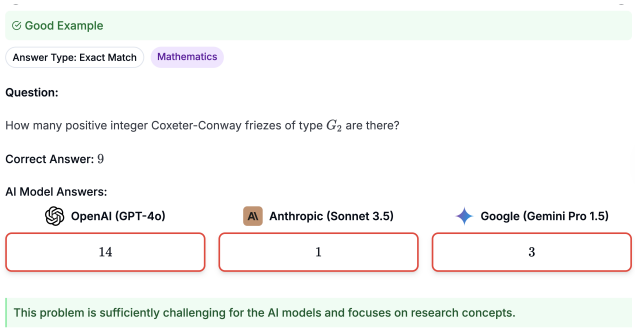
Process

1. **Write Question.** You will first write a valid, very difficult question in English that AIs and average humans cannot answer. We aim to get questions that only exceptional individuals can answer correctly, so try and create questions that would impress you if an AI solved them. Again, please write questions in English.

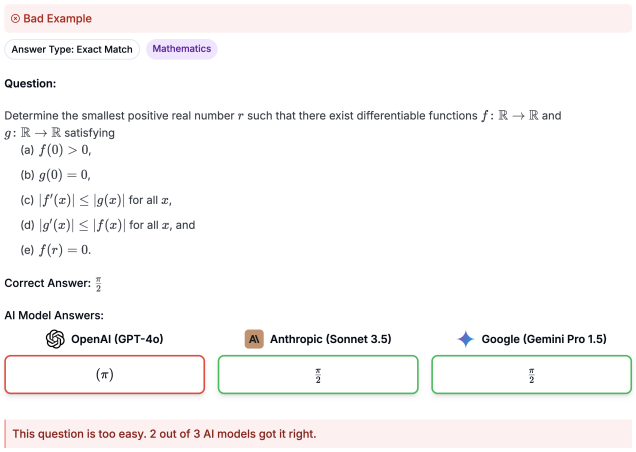
2. **AIs Assess Question Difficulty.** Upon creating a question, we will ask state-of-the-art AI models to answer the question to help determine whether the question is too easy.

3. **Write Answer Explanation.** If the question is not too easy for the AIs, then we will ask you to write a thorough but concise solution for your question.

4. **Peer Review.** After submitting your question, answer, and rationale, your response will be saved, and we will conduct another round of manual review to maintain the quality of the benchmark. The answer rationale will help experts and AIs determine whether the answer you provided is



Good Example



Bad Example

Supplementary Information Fig 1: Examples of good and bad questions. We show question contributors on our submission platform a few good and bad examples as part of the guidelines.

correct. Authors can make changes or remove their submitted questions any time at the dashboard.

5. **Publish.** If your question is selected for the dataset, your contribution will be highlighted alongside the question. People who write more accepted questions will appear earlier in the author list of the paper. (A small fraction of these questions will be kept private to detect if an AI is cheating and memorizing answers. You will still be credited in the paper if your question is put in the private cheating detection dataset. Coauthorship on the paper is optional.)

Guidelines

Original. Questions must be original, not copy and pasted from someone else. You may possibly include questions from your exams that you have created, provided the answers are not publicly available and the questions are extremely challenging. Do not submit multiple questions that are minor variants of each other or numerous similar questions of the same theme.

Challenging. Questions should be difficult for non-experts and not easily answerable by everyday people. Questions should not be simple trick questions. Questions should not be straightforward calculation/computation questions. The questions for this dataset can be so challenging that they are impractical for real-world human exams or problem sets. The answers should not be easily found through a Google search. It should impress you if an AI correctly answers your question. We found questions written by undergraduates tend to be too easy for the models. As a rough rule of thumb, if a randomly selected undergraduate can understand what is being asked, it is likely too easy for the frontier LLMs of today and tomorrow. Avoid using images if the question can easily be expressed in text. Examples of good and bad submissions are shown in Supplementary Information Fig 1.

Objective, Self-Contained, Close-Ended. Questions should have answers accepted by other experts with relevant expertise. No highly disputed, "personal taste," ambiguous, or subjective questions. All necessary context and definitions should be included within the question (do not link to outside sources). It is acceptable to use technical jargon or notation without definition as long as it is standard, unambiguous, and would be widely recognized within your specialty. Answers should not should not have many decimals or be imprecise. Exact match questions need to be closed-ended with specific, univocal and succinct answers. Questions asking to "explain", "discuss", or "describe" are likely unsuitable for the exact match format.

No weapon questions. No questions about virology. No questions highly related to chemical, biological, radiological, nuclear weapons, or cyberweapons used for attacking critical infrastructure.

2 Human Review Instructions

Questions which merely stump models are not necessarily high quality – they could simply be adversarial to models without testing advanced knowledge. To resolve this, we employ two rounds of human reviews to ensure our dataset is thorough and sufficiently challenging as determined by human experts in their respective domains. The primary goal is to help the question contributors (who are primarily academics and researchers from a wide range of disciplines)

better design questions that are closed-ended, robust, and of high quality for AI evaluation.

2.1 Review Round 1

We recruit human subject expert reviewers to score, provide feedback, and iteratively refine all user submitted questions. This is similar to the peer review process in academic research, where reviewers give feedback to help question submitters create better questions. We train all reviewers on the instructions and rubric below.

Reviewer Instructions

- Questions should usually (but do not always need to) be at a graduate / PhD level or above. (Score 0 if the question is not complex enough and AI models can answer it correctly.)
  - If the model is not able to answer correctly and the question is below a graduate level, the question can be acceptable.
- Questions can be any field across STEM, law, history, psychology, philosophy, trivia, etc. as long as they are tough and interesting questions.
  - For fields like psychology, philosophy, etc. we usually check if the rationale contains some reference to a book, paper or standard theories.
  - For fields like law, the question text can be adjusted with "as of 2024". Make sure questions about law are time-bounded.
  - Questions do not always need to be academic. A handful of movie, TV trivia, classics, history, art, or riddle questions in the dataset are OK.
  - Trivia or complicated game strategy about chess, go, etc. are okay as long as they are difficult.
  - We generally want things that require a high level of human intelligence to figure out.
- Questions should ask for something precise and have an objectively correct, univocal answer.
  - If there is some non-standard jargon for the topic/field, it needs to be explained.
  - Questions must have answers that are known or solvable.
  - Questions should not be subjective or have personal interpretation.
  - Questions like "Give a proof of..."; "Explain why..."; "Provide a theory that explains..." are usually bad because they are not closed-ended and we cannot evaluate them properly. (Score 0)

  - No questions about morality or what is ethical/unethical. (Score 0)

- Questions should be original and not derived from textbooks or Google. (Score 0 if searchable on web)
- Questions need to be in English. (Score 1 and ask for translation in the review if the question is written in a different language)
- Questions should be formatted properly. (Score 1–3 depending on degree of revisions needed)
  - Question with numerical answers should have results approximated to max 2–3 decimals.
  - Fix LaTeX formatting if possible. Models often get questions right after LaTeX formatting is added or improved.
  - Questions that can be converted to text should be (converting images to text often helps models get them right).

Other Tips

- Please write detailed justifications and feedback. This is going out to the question submitter so please use proper language and be respectful.
  - Explanations should include at least some details or reference. If the rationale is unclear or not detailed, ask in the review to expand a bit.
  - Please check if the answer makes sense as a possible response to the question, but if you do not have knowledge/context, or if it would take more than 5 minutes to solve, that is okay.
- Please prioritize questions with no reviews and skip all questions with more than 3 reviews.
- Please double check that the model did actually answer the question wrong.
  - Sometimes the exact match feature does not work well enough, and there are false negatives. We have to discard any exact match questions that a model got right.
- On the HLE dashboard, look at least 10 examples reviewed by the organizers before starting to review, and review the examples from training.
- The average time estimated to review a question 3–5 minutes.
- Use a "-1 Unsure" review if the person submitting seems suspicious or if you're not convinced their answer is right.

Review Round 1 scoring rubric Reviewers assign one of the following scores based on the overall quality, originality, and difficulty of the question:

Score 0 – Discard The question is out of scope, not original, spam, or otherwise not good enough to be included in the HLE set and should be discarded.

Score 1 – Major Revisions Needed Major revisions are needed for this question, or the question is too easy and simple.

Score 2 – Some Revisions Needed The difficulty and expertise required to answer the question are borderline. Some revisions are needed for this question.

Score 3 – Okay The question is sufficiently challenging but the knowledge required is not graduate-level nor complex. Minor revisions may be needed for this question.

Score 4 – Great The knowledge required is at the graduate level, or the question is sufficiently challenging.

Score 5 – Top-Notch The question is top-notch and essentially perfect.

Unsure The reviewer is unsure if the question fits the HLE guidelines, or unsure if the answer is right.

2.2 Review Round 2

To thoroughly refine our dataset, we train a set of reviewers along with organizers to pick the best questions. These reviewers are identified by organizers from Round 1 reviews as particularly high quality and thorough in their feedback. Different than the first round of reviews, reviewers are asked to grade both the question and look at feedback from Round 1 reviewers. Organizers then approve questions based on reviewer feedback in this round. We employ a new rubric for this round below.

Review Round 2 scoring rubric In Round 2, reviewers select one of the following scores, taking into account both the question itself and Round 1 feedback:

Score 0 – Discard The question is out of scope, not original, spam, or otherwise not good enough to be included in the HLE set and should be discarded.

Score 1 – Not sure Major revisions are needed for this question, or the reviewer is unsure about the question. The reviewer should put their thoughts in the comment box for an organizer to evaluate.

Score 2 – Pending The reviewer believes there are still minor revisions needed on this question. Comments should be provided and an organizer will make the final decision.

Score 3 – Easy questions models got wrong These are very basic questions that models got correct or the question was easily found online. Any questions which are artificially difficult (e.g., large calculations needing a calculator, requires running/rendering code, etc.) should also belong in this category, since the models we evaluate cannot access these tools and this creates artificial difficulty. "Found online" here means easily discoverable via a simple web search; research papers, journals, and books are acceptable sources.

Score 4 – Borderline The question is not interesting, or the question is sufficiently challenging but one or more of the models got the answer correct.

Score 5 – Okay to include in HLE benchmark Very good questions (often those which had a score of 3 in the previous review round). The reviewer believes it should be included in the HLE benchmark.

Score 6 – Top question in its category Great questions (often those which had a score of 4–5 in the previous review round), at a graduate or research level. For non-STEM questions and trivia, "graduate level" is interpreted more loosely; they are acceptable as long as they are challenging and interesting.